

ON RATES OF CONVERGENCE FOR THE OVERLAP IN THE HOPFIELD MODEL

PETER EICHELSBACHER¹ AND BASTIAN MARTSCHINK²

Abstract: We consider the Hopfield model with n neurons and an increasing number $p = p(n)$ of randomly chosen patterns and use Stein's method to obtain rates of convergence for the central limit theorem of overlap parameters, which holds for every fixed choice of the overlap parameter for almost all realisations of the random patterns.

AMS 2000 Subject Classification: Primary 60F05; Secondary 82B20, 82B26.

Key words: Stein's method, exchangeable pairs, Hopfield model, overlap, neural networks.

1. INTRODUCTION

1.1 The Hopfield model The so-called Hopfield model was introduced by Figotin and Pastur in [15] and [16] as a model for a spin glass. They studied a class of spin glass models which also included the one with the energy function known today as the Hopfield model, which was also introduced by Hopfield in [14] in the context of neural networks as a model for an associative memory with $n \in \mathbb{N}$ neurons. Thus Hopfield linked the study of neural networks to the one of spin models. The success of this model was mainly based on this reinterpretation of the model and therefore it may be right to call it the Hopfield model. Being a model for the associate (also termed content-addressable) memory it is not derived directly from a physical or biological system. Roughly speaking, the recognition and/or retrieval of one out of $p \in \mathbb{N}$ stored patterns constitutes the central problem of the model. This means that one wants to store a certain amount of information and perform the quite difficult task to recognize it on the basis of partial or corrupted data, which is not easy for a usual search algorithm.

We consider a system of $n \in \mathbb{N}$ neurons. Each neuron can be in one of two possible states, either -1 or 1 . We will denote by $\sigma_i \in \{-1, 1\}$ the neural activity of the i^{th} neuron, $i \in \{1, \dots, n\}$

¹Ruhr-Universität Bochum, Fakultät für Mathematik, NA 3/66, D-44780 Bochum, Germany, peter.eichelsbacher@rub.de

²Hochschule Bonn-Rhein Sieg, Fachbereich 03, B 295, D-53757 Sankt Augustin, Germany, bastian.martschink@h-brs.de

The authors have been supported by Deutsche Forschungsgemeinschaft via SFB/TR 12.

and thus, in the context of spin systems, σ_i would be the spin variable at $i \in \{1, \dots, n\}$. Thus a spin configuration $(\sigma_1, \dots, \sigma_n)$ is taken from the set of spin configurations $\{-1, 1\}^n$. In general the instantaneous configuration of all the spin variables at a given time describes the state of such a network. Furthermore let $(\Omega, \mathcal{B}, \mathbb{P})$ be an abstract probability space. The model consists of $p \in \mathbb{N}$ stored patterns on this space which will be denoted by ξ^μ , $\mu \in \{1, \dots, p\}$. Thus $\xi^\mu = (\xi_1^\mu, \dots, \xi_n^\mu) \in \{-1, 1\}^n$ describes the codification of the μ^{th} stored pattern. $(\sigma_i)_{i \in \mathbb{N}}$ and $(\xi_i^\mu)_{i \in \mathbb{N}}$ with $\mu \in \mathbb{N}$ are considered to be random variables and we will assume that the family of random variables $\{\sigma_i, \xi_j^\mu \mid i, j, \mu \in \mathbb{N}\}$ is independent. Additionally we assume that the random variables satisfy $\mathbb{P}(\sigma_i = \pm 1) = 1/2$ and $\mathbb{P}(\xi_j^\mu = \pm 1) = 1/2$. Thus we denote by $\mathbb{P}_\xi = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1)^{\otimes \mathbb{N}^2}$ the marginal distribution of the patterns $\xi = (\xi_i^\mu)_{i, \mu \in \mathbb{N}}$, and similarly, by $\mathbb{P}_\sigma = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1)^{\otimes \mathbb{N}}$ the marginal distribution of the spin variables $\sigma = (\sigma_i)_{i \in \mathbb{N}}$. As $n \rightarrow \infty$ p can either be fixed or increasing with n . Now let

$$H_n(\sigma, \xi) = -\frac{1}{2n} \sum_{\mu=1}^p \sum_{i,j=1}^n \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j, \quad n \in \mathbb{N}, \quad (1.1)$$

denote the Hopfield Hamiltonian. At this point one might notice the spin-flip dynamic $H_n(-\sigma, \xi) = H_n(\sigma, \xi)$, showing that the Hopfield model cannot distinguish between a spin configuration and its negative. Governed by this Hamiltonian, [1] presented a generalized Glauber single-spin dynamics on the set of spin configurations at finite temperature $1/\beta \in (0, \infty)$, which describes a reversible and irreducible Markov process. The equilibrium distribution of this process is the finite-volume Gibbs measure

$$dP_{n,\beta,\xi}(\sigma) = \frac{1}{Z_{n,\beta,\xi}} \exp(-\beta H_n(\sigma, \xi)) d\mathbb{P}_\sigma, \quad (1.2)$$

where the partition function $Z_{n,\beta,\xi}$ is the appropriate normalization.

In the sequel the focus of attention will be on the investigation of the behavior of the so-called overlap under the equilibrium distribution $P_{n,\beta,\xi}$ as $n \rightarrow \infty$. Let

$$\xi_i = (\xi_i^\mu)_{\mu \in \{1, \dots, p\}}, \quad i \in \{1, \dots, n\}, \quad (1.3)$$

be the vector consisting of the i^{th} components of the first p patterns. If p is not constant and grows with n , $\xi_i \in \mathbb{R}^p$ still depends on n via the dimension. We define the *overlap* by

$$\frac{1}{n} S_n(\sigma, \xi) = \frac{1}{n} \sum_{i=1}^n \xi_i \sigma_i \in \mathbb{R}^p, \quad (1.4)$$

with $\xi_i \sigma_i = (\xi_i^1 \sigma_i, \dots, \xi_i^p \sigma_i)^t$. With the overlap we obtain a comparison between the spin configuration σ and the stored patterns ξ^μ , $\mu \in \{1, \dots, p\}$, meaning that the μ^{th} overlap parameter - the μ^{th} component of (1.4) - equals one if and only if $\sigma_i = \xi_i^\mu$ for all $i \in \{1, \dots, n\}$. Definition (1.4) provides the opportunity to express the Hamiltonian (1.1) in a more convenient way. It can be rewritten as the quadratic function of the overlap

$$H_n(\sigma, \xi) = -\frac{n}{2} \left\| \frac{1}{n} S_n(\sigma, \xi) \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^p . If there is no opportunity for confusion we will drop the explicit dependence on σ and ξ and write S_n and H_n instead of $S_n(\sigma, \xi)$ and $H_n(\sigma, \xi)$, respectively.

In the case $p = 1$ the Hopfield model and the Curie-Weiss model are the same apart from a change of variable. The Curie-Weiss model is a well-known approximation of the Ising-model.

The classical theory of magnetism occupies a central place in the physical literature. It allows the study of the behavior of thermodynamic quantities such as the specific heat, isothermal susceptibility, and magnetization in the neighborhood of the critical point. Because of its relative simplicity and the qualitative correctness of at least some of its predictions, it has been historically important. For our investigation of the Hopfield model we focus on the so called Curie-Weiss equation given by

$$\beta x = \operatorname{arctanh}(x). \quad (1.5)$$

This equation is also called *mean field* or *fixed point* equation. Its derivation can for example be found in [10]. Of course this equation may have many solutions. Let $x^\pm(\beta)$ denote for $\beta > 0$ the largest (respectively smallest) solution $x \in (-1, 1)$ of (1.5). It was shown that $x^+(\beta) = -x^-(\beta) \neq 0$ for $\beta > \beta_c$, where $\beta_c = 1$ is the critical inverse temperature. For $\beta \leq \beta_c$ we have $x^\pm(\beta) = 0$. This definition of the Curie-Weiss equation can be extended to the case of the external magnetic field with strength $h \neq 0$ yielding

$$\beta x + h = \operatorname{arctanh}(x). \quad (1.6)$$

Here let $x(\beta, h)$ denote the solution of (1.6) which satisfies $\operatorname{sign}(x) = \operatorname{sign}(h)$. As we will see these solutions of the Curie-Weiss equation discussed above play an important role when discussing the Hopfield model. Abbreviate

$$x^* := \begin{cases} x^+(\beta), & \text{if } h=0, \\ x(\beta, h), & \text{otherwise.} \end{cases}$$

For investigating the behaviour of the overlap, we also extend the notion of the Gibbs measure $P_{n,\beta,\xi}$ given in (1.2) to the case of an external magnetic field he_l with strength $h \neq 0$ in the direction of the l^{th} unit vector $e_l \in \mathbb{R}^p$. Thus, let

$$dP_{n,\beta,he_l,\xi}(\sigma) = \frac{1}{Z_{n,\beta,he_l,\xi}} \exp(-\beta H_n + \langle S_n, he_l \rangle) d\mathbb{P}_\sigma, \quad (1.7)$$

where $Z_{n,\beta,he_l,\xi}$ denotes the appropriate normalization.

For $\beta > 0$ and $h \neq 0$ having the direction of the l^{th} unit vector e_l it was shown in [4] that for \mathbb{P}_ξ -almost all realizations of the patterns ξ and if $p/n \rightarrow 0$ the overlap $\frac{S_n}{n}$ satisfies the *law of large numbers*

$$P_{n,\beta,he_l,\xi} \left(\frac{S_n}{n} \in d\nu \right) \Rightarrow \delta_{\pm x(\beta,h)e_l}(d\nu) \text{ as } n \rightarrow \infty.$$

The authors in [4] stated that the condition on p is the weakest possible under which the law of large numbers is satisfied. Note that for $\beta \leq \beta_c = 1$ we have $x(\beta, h) = 0$ and thus δ_0 is the unique limiting measure in the high-temperature region. For $\beta > 1$ it was mentioned that the measures of the law of large numbers are all distinct and they were referred to as so-called *extremal measures*.

The corresponding *large deviation principle* (LDP for short) was established in [2]. Under the assumption $p(n)/n \rightarrow 0$ for almost all ξ the sequence $(\frac{S_n}{n})_n$ under the Gibbs measure $P_{n,\beta,\xi}$ obeys a LDP with speed n and *deterministic* rate function I . If the inverse temperature β is different from the critical inverse temperature $\beta_c = 1$ and $p(n)/n \rightarrow \infty$, the overlap parameter multiplied by n^γ with $1/2 < \gamma < 1$ obeys a LDP with speed $n^{1-\gamma}$ and a quadratic rate function, see [7]. The latter result is known as a moderate deviations principle (MDP for short).

On the scale of fluctuations, when analysing the distribution of $\sqrt{n}(S_n/n - x^*e_l)$, the disorder becomes visible. Indeed, for $p(n)/n \rightarrow 0$ and $(\beta, h) \neq (1, 0)$ the overlap under $P_{n,\beta,\xi}$ satisfies P_ξ -almost surely a central limit theorem with a covariance matrix which could be expected from the analogy with the Curie-Weiss model and a centering which differs in the case $\beta > 0$ or $h \neq 0$ from the naively expected one by a ξ -dependent adjustment, see [11] and [3]. In this paper we are aiming to give an alternative proof of these central limit theorems for the overlap parameter under $P_{n,\beta,\xi}$. We will apply Stein's method. This method has emerged as a powerful tool for assessing the quality of distributional approximations and it is notable for avoiding the use of transforms, and for supplying bounds, such as those of Berry-Esseen quality, on approximation error in the presence of dependence. We will be able to present rates of convergence for central limit theorems for the overlap parameter, which are optimal for the Hopfield model with a finite number of randomly chosen patterns. As in the Curie-Weiss model at the critical temperature $(\beta, h) = (1, 0)$ the fluctuations are non Gaussian and the limiting distribution has a random component, see [13] and [23]. Interesting enough the random term occurring in the central limit theorem is no longer present on a moderate deviations scale, where the overlap parameter has to be multiplied by n^γ with $1/4 < \gamma < 1$: here for certain choices of $p(n)$ the rescaled overlap parameter obeys a MDP with speed $n^{1-4\gamma}$ and a rate function that is basically a fourth power, see [7]. Anyhow, in this paper we do not consider the case $(\beta, h) = (1, 0)$.

1.2 Statement of the main results

General assumption. From now on we make the assumption that $p = p(n)$, $p \leq n$ is a nondecreasing function of n for all $n \in \mathbb{N}$.

As in [12] we choose a preferred pattern in two different ways. We consider the unbiased Hamiltonian (1.1) and investigate the fluctuations under the condition that the overlap is already in a neighbourhood of x^*e_l . Alternatively, the preferred pattern can be chosen by introducing the magnetic field as in (1.7). In the case of (1.1) with $\beta < \beta_c$ the central limit theorem holds with center zero. Otherwise the limit theorem requires a ξ -dependent adjustment of a deterministic centering. Therefore one has to control the influence of the random patterns. For fixed $\epsilon > 0$ we define

$$\begin{aligned} \alpha &:= \frac{1}{n} \max \left\{ p, \left(\frac{3 \log n}{\log(1 + \epsilon)} \right)^4 \right\}, \\ \epsilon_n &:= \sqrt{\alpha}(2 + \sqrt{\alpha})(1 + \epsilon). \end{aligned} \tag{1.8}$$

By [12, Proposition 2.1] we see that the operator norm of $\Sigma^n(\xi) = \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^t - \text{Id}_{\mathbb{R}^p}$ converges to zero for P_ξ -almost all ξ : for P_ξ -almost all ξ , there exists an $n_0(\xi) \in \mathbb{N}$ such that for all $n \geq n_0(\xi)$

$$\|\Sigma^n(\xi)\| \leq \epsilon_n. \tag{1.9}$$

The following index set depends on the dimension p , on the inverse temperature β , the presence or absence of an external magnetic field h and its direction e_l :

$$L := \begin{cases} \{\text{sign}(h)l\}, & \text{in the case } h \neq 0, \\ \{1\}, & \text{in the case } 0 < \beta < \beta_c \text{ and } h = 0, \\ \{-p, \dots, -1, 1, \dots, p\}, & \text{in the case } \beta > \beta_c \text{ and } h = 0. \end{cases} \tag{1.10}$$

The index set L is used to describe those directions that the overlap favors under the equilibrium measure. In $(\beta_c, 0)$ the central limit theorem fails (see [12]). Thus we do not need L for these

parameters. The following result is proved in [12, Proposition 2.3] and is an important step for defining the centering.

Proposition 1.1.

Let $\beta > 0$ and $h \geq 0$ such that $(\beta, h) \neq (\beta_c, 0)$ and $l \in \{-p, \dots, -1, 1, \dots, p\}$. For $\lambda \in \mathbb{R}^p$, we define the ξ -dependent function

$$\Phi(\lambda) := -\frac{1}{2\beta} \|\lambda - h e_l\|^2 + \frac{1}{n} \sum_{j=1}^n \log \cosh \langle \lambda, \xi_j \rangle. \quad (1.11)$$

Then, for all strictly positive $c_1 < (1 - \beta(1 - (x^*)^2))/\beta$, there exists an $r_1 > 0$, depending on β , h and c_1 only, and for \mathbb{P}_ξ -almost all ξ , there exists an $n_1(\xi) \geq n_0(\xi)$, which does not depend on the choice of l , such that for all $n \geq n_1(\xi)$ the following assertions hold:

- (1) For all λ in the closed ball $\overline{B_{r_1}(\text{arctanh}(x^* e_l))}$, the matrix $-D^2\Phi(\lambda)$ is uniformly positive definite in the sense that

$$\langle u, -D^2\Phi(\lambda)u \rangle \geq c_1 \|u\|^2 \text{ for all } u \in \mathbb{R}^p.$$

- (2) On the set $\overline{B_{r_1}(\text{arctanh}(x^* e_l))}$, the map Φ has a unique maximum which is attained in the point $\lambda_l^n(\xi)$ satisfying

$$|\lambda_l^n(\xi) - \text{arctanh}(x^* e_l)| \leq c_2 \epsilon_n$$

with $c_2 = 2|x|/c_1$. In particular, $\lambda_l^n(\xi) = 0$ in the case $\beta < \beta_c$ and $h = 0$.

Remark 1.2. The function Φ defined in (1.11) is sometimes called *quenched free-energy* of the Hopfield model. If the realizations ξ_1, \dots, ξ_n take all possible values with the same frequency and n is a multiple of 2^p , then $\lambda_l^n(\xi) = \text{arctanh}(x^* e_l)$.

The random centering is given by

$$x_l^n(\xi) = \frac{1}{\beta} (\lambda_l^n(\xi) - h e_l) \quad (1.12)$$

with the help of $\lambda_l^n(\xi)$ for $l \in \{-p, \dots, -1, 1, \dots, p\}$. Even if it is not indicated by the name it remains important to notice that (1.12) still depends on β and h . We have to extend this definition because (1.12) is only defined for \mathbb{P}_ξ -almost all ξ and $n \geq n_1(\xi)$. We assign

$$x_l^n(\xi) = \frac{1}{\beta} (\text{arctanh } x^* - h) e_l = x^* e_l \quad (1.13)$$

whenever $\lambda_l^n(\xi)$ is not defined. The second equality of (1.13) is due to the Curie-Weiss equation (1.6). Using Proposition 1.1 we see that for $\beta < \beta_c$ the centering satisfies $x_l^n(\xi) = 0$, while for $\beta > \beta_c$ the centering is close to the limiting point x^* in the sense that

$$\|x_l^n(\xi) - x^* e_l\| \leq \frac{1}{\beta} c_2 \epsilon_n \rightarrow 0 \quad (1.14)$$

as $n \rightarrow \infty$ for some constant C and ϵ_n defined in (1.8).

From now on we will write random vectors in \mathbb{R}^d in the form $w = (w_1, \dots, w_d)^t$, where w_i are \mathbb{R} -valued variables for $i = 1, \dots, d$. If a matrix Σ is symmetric, nonnegative definite, we denote by $\Sigma^{1/2}$ the unique symmetric, nonnegative definite square root of Σ . Id denotes the identity matrix and from now on Z will denote a random vector having standard multivariate normal distribution. The expectation with respect to the measure $P_{n,\beta,h e_l,\xi}$ will be denoted by $\mathbb{E} := \mathbb{E}_{P_{n,\beta,h e_l,\xi}}$.

Let $\pi_k : \mathbb{R}^p \rightarrow \mathbb{R}^k$ (with $k \leq p$) denote the canonical projection.

Theorem 1.3.

Let $\beta, h > 0$, $l \in \mathbb{Z}$, $l \neq 0$, and $k \in \mathbb{N}$. We assume that p depends on n in a nondecreasing way satisfying $p \leq n$. Let $x = x_l^n(\xi)$ be defined as in (1.12) and W be the following random variable:

$$W := \sqrt{n} \pi_k \left(\frac{S_n}{n} - x \right).$$

If Z has the k -dimensional standard normal distribution, under the measure $P_{n,\beta,h\epsilon_l,\xi}$, we have, for every three times differentiable function g and \mathbb{P}_ξ -almost all ξ ,

$$\left| \mathbb{E}g(W) - \mathbb{E}g(\Sigma^{1/2}Z) \right| \leq C \max \left\{ p\sqrt{p}\epsilon_n, \frac{p^2}{n^{1/2}} \right\},$$

for a constant C and $\Sigma := \mathbb{E}[W W^t]$.

Remark 1.4. The rate of convergence obtained here is useless unless

$$\max \left\{ p\sqrt{p}\epsilon_n, \frac{p^2}{n^{1/2}} \right\} \rightarrow 0. \quad (1.15)$$

In [3, Theorem 1.1] the authors proved that the condition $p/n \rightarrow 0$ is sufficient in order to state the central limit theorem and show the weak convergence. In [12] and [3] there is no information available on the speed of convergence. Obviously (1.15) is poorer but we do not need any conditions on p in advance. Our theorem implies weak convergence.

In order to state a result for non-smooth test functions g in the multivariate setting, we introduce a class of test functions \mathcal{G} following [19]. Let again Φ denote the standard normal distribution function in \mathbb{R}^d . We define for $g : \mathbb{R}^d \rightarrow \mathbb{R}$

$$g_\delta^+(x) = \sup\{g(x+y) : |y| \leq \delta\}, \quad (1.16)$$

$$g_\delta^-(x) = \inf\{g(x+y) : |y| \leq \delta\}, \quad (1.17)$$

$$\tilde{g}(x, \delta) = g_\delta^+(x) - g_\delta^-(x). \quad (1.18)$$

Let \mathcal{G} be a class of real measurable functions on \mathbb{R}^d such that

- (1) The functions $g \in \mathcal{G}$ are uniformly bounded in absolute value by a constant, which we take to be 1 without loss of generality.
- (2) For any $d \times d$ matrix A and any vector $b \in \mathbb{R}^d$, $g(Ax + b) \in \mathcal{G}$.
- (3) For any $\delta > 0$ and any $g \in \mathcal{G}$, $g_\delta^+(x)$ and $g_\delta^-(x)$ are in \mathcal{G} .
- (4) For some constant $a = a(\mathcal{G}, d)$, $\sup_{g \in \mathcal{G}} \left\{ \int_{\mathbb{R}^d} \tilde{g}(x, \delta) \Phi(dx) \right\} \leq a\delta$.

Obviously we may assume $a \geq 1$. Considering the one dimensional case, we notice that the collection of indicators of all half lines and indicators of all intervals form classes in \mathcal{G} that satisfy these conditions with $a = \sqrt{2/\pi}$ and $a = 2\sqrt{2/\pi}$ respectively. This was shown for example in [18]. In dimension $d \geq 1$ the class of indicators of convex sets is known to be such a class. Using this class of functions we are able to present rates of convergence for non-smooth test functions.

Theorem 1.5.

Let $\beta, h > 0$, $l \in \mathbb{Z}$, ($l \neq 0$) and $k \in \mathbb{N}$. We assume that p depends on n in a nondecreasing way satisfying $p \leq n$. Let $x = x_l^n(\xi)$ be defined as in (1.12) and W be as in Theorem 1.3. If Z has the k -dimensional standard normal distribution, under the measure $P_{n,\beta,he_l,\xi}$, we have, for all $g \in \mathcal{G}$ with $|g| \leq 1$ and \mathbb{P}_ξ -almost all ξ ,

$$\left| \mathbb{E}g(W) - \mathbb{E}g\left(\Sigma^{1/2}Z\right) \right| \leq C \log(n) \max \left\{ p\sqrt{p}\epsilon_n, \frac{p^2}{n^{1/2}} \right\},$$

for a constant C and $\Sigma := \mathbb{E}[W W^t]$.

In the case where p is fixed the rate gets much simpler since we do not need the projection in order to reduce the size of the vector W .

Theorem 1.6.

Let $\beta, h > 0$, $l \in \mathbb{Z}$ and $l \neq 0$. We assume that p is fixed. Let $x = x_l^n(\xi)$ be defined as in (1.12) and W be the following random variable:

$$W := \sqrt{n} \left(\frac{S_n}{n} - x \right).$$

If Z has the p -dimensional standard normal distribution, under the measure $P_{n,\beta,he_l,\xi}$, we have, for every three times differentiable function g and \mathbb{P}_ξ -almost all ξ ,

$$\left| \mathbb{E}g(W) - \mathbb{E}g\left(\Sigma^{1/2}Z\right) \right| \leq C n^{-1/2},$$

for a constant C and $\Sigma := \mathbb{E}[W W^t]$.

With the same techniques necessary to prove Theorem 1.5 we get a theorem similar to Theorem 1.6 with rate $\log(n)n^{-1/2}$.

When there is no external field it is natural to ask for the fluctuations of the overlap around x^*e_l . With L as in (1.10) we determine the conditional fluctuations and a rate of convergence:

Theorem 1.7.

Let $\beta > 0$, $\beta \neq \beta_c$, $h = 0$, $l \in L$ and $k \in \mathbb{N}$. We assume that p depends on n in a nondecreasing way satisfying $p \leq n$. Let $x = x_l^n(\xi)$ be defined as in (1.12) and W be as in Theorem 1.3. Then, if Z has the k -dimensional standard normal distribution, under the conditional measure

$$P_{n,\beta,\xi} \left(\cdot \mid \frac{S_n}{n} \in B(x^*e_l, \epsilon) \right),$$

we have for every three times differentiable function g and \mathbb{P}_ξ -almost all ξ ,

$$\left| \mathbb{E}g(W) - \mathbb{E}g\left(\Sigma^{1/2}Z\right) \right| \leq C \max \left\{ p\sqrt{p}\epsilon_n, \frac{p^2}{n^{1/2}} \right\},$$

for a constant C and $\Sigma := \mathbb{E}[W W^t]$.

Note that also for the case of $h = 0$ a theorem for non-smooth test functions could be stated, similar to Theorem 1.5, and additionally we obtain a theorem if p is fixed with rate $n^{-1/2}$ in the same way as in Theorem 1.6.

In Section 2 of the present paper, we introduce Stein's method and present two plug-in theorems for multivariate normal approximation. Section 3 contains some auxiliary results which will be necessary for the proofs given in Section 4.

2. STEIN'S METHOD OF EXCHANGEABLE PAIRS

Starting with a bound for the distance between univariate random variables and the normal distribution Stein's method was first published in [20] (1972). In [21] Stein introduced his exchangeable pair approach. At the heart of the method is a coupling of a random variable W with another random variable W' such that (W, W') is *exchangeable*, i.e. their joint distribution is symmetric. Stein proved further on that a measure of proximity of W to normality may be provided by the exchangeable pair if $W' - W$ is sufficiently small. He assumed the property that there is a number $\lambda > 0$ such that the expectation of $W' - W$ with respect to W satisfies

$$\mathbb{E}[W' - W|W] = -\lambda W.$$

Heuristically, this condition can be understood as a linear regression condition: if (W, W') were bivariate normal with correlation ϱ , then $\mathbb{E}[W'|W] = \varrho W$ and the condition would be satisfied with $\lambda = 1 - \varrho$. Stein proved that for any uniformly Lipschitz function h

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \leq \delta \|h'\|$$

with Z denoting a standard normally distributed random variable and

$$\delta = 4\mathbb{E}\left|1 - \frac{1}{2\lambda}\mathbb{E}[(W' - W)^2|W]\right| + \frac{1}{2\lambda}\mathbb{E}|W - W'|^3.$$

Stein's approach has been successfully applied in many models, see e.g. [21] or [22] and references therein. In [18] the range of application was extended by replacing the linear regression property by a weaker condition assuming that there is also a random variable $R = R(W)$ such that

$$\mathbb{E}[W' - W|W] = -\lambda W + R.$$

While the approach has proved successful also in non-normal contexts (see [5],[6] and [8]) it remained restricted to the one-dimensional setting for a long time. Applying the linear regression heuristic in the multivariate case leads to a new condition due to [17]:

$$\mathbb{E}[W' - W|W] = -\Lambda W + R \tag{2.1}$$

for an invertible $d \times d$ matrix Λ and a remainder term $R = R(W)$. Different exchangeable pairs, obviously, will yield different Λ and R .

The theorems for smooth test functions are based on a nonsingular multivariate normal approximation theorem taken from [17]. To present this theorem we fix some more notations. The transpose of the inverse of a matrix will be presented in the form $A^{-t} := (A^{-1})^t$. Furthermore we will need the supremum norm, denoted by $\|\cdot\|$ for both functions and matrices. For derivatives of smooth functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we use the notation ∇ for the gradient operator. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we abbreviate

$$|f|_1 := \sup_i \left\| \frac{\partial}{\partial x_i} f \right\|, \quad |f|_2 := \sup_{i,j} \left\| \frac{\partial^2}{\partial x_i \partial x_j} f \right\|,$$

and so on, if these derivatives exist.

Theorem 2.1. *(Reinert, Röllin: 2009)*

Assume that (W, W') is an exchangeable pair of \mathbb{R}^d -valued random vectors such that

$$\mathbb{E}[W] = 0, \quad \mathbb{E}[W W^t] = \Sigma,$$

with $\Sigma \in \mathbb{R}^{d \times d}$ symmetric and positive definite. If (W, W') satisfies (2.1) for an invertible matrix Λ and a $\sigma(W)$ -measurable random vector R and if Z has d -dimensional standard normal distribution, we have for every three times differentiable function g ,

$$\left| \mathbb{E}g(W) - \mathbb{E}g(\Sigma^{1/2}Z) \right| \leq \frac{|g|_2}{4}A + \frac{|g|_3}{12}B + \left(|g|_1 + \frac{1}{2}d\|\Sigma\|^{1/2}|g|_2 \right) C, \quad (2.2)$$

where, with $\lambda^{(i)} := \sum_{m=1}^d |(\Lambda^{-1})_{m,i}|$,

$$\begin{aligned} A &= \sum_{i,j=1}^d \lambda^{(i)} \sqrt{\mathbb{V}[\mathbb{E}[(W'_i - W_i)(W'_j - W_j) \mid W]]}, \\ B &= \sum_{i,j,k=1}^d \lambda^{(i)} \mathbb{E}|(W'_i - W_i)(W'_j - W_j)(W'_k - W_k)|, \\ C &= \sum_{i=1}^d \lambda^{(i)} \sqrt{\mathbb{V}[R_i]}. \end{aligned}$$

The advantage of Stein's method is that the bounds to a multivariate normal distribution reduce to the computation of, or bounds on, low order moments, here bounds on the absolute third moments, on a conditional variance and on the variance of the remainder term. Such variance computations may be difficult, but we will get rates of convergence at the same time. In the same context as in [17] the authors in [9] proved the following theorem, presenting bounds for non smooth test functions. Their development differs from [17] using the relationship to the bounds in [18].

Theorem 2.2.

Let (W, W') be an exchangeable pair with $\mathbb{E}[W] = 0$ and $\mathbb{E}[W W^t] = \Sigma$ with $\Sigma \in \mathbb{R}^{d \times d}$ symmetric and positive definite. Again we assume that (W, W') satisfies (2.1) for an invertible matrix Λ and a $\sigma(W)$ -measurable random vector R and additionally, for $i \in \{1, \dots, d\}$, $|W'_i - W_i| \leq A$. Then,

$$\begin{aligned} \sup_{g \in \mathcal{G}} |\mathbb{E}g(W) - \mathbb{E}g(\Sigma^{1/2}Z)| &\leq C \left[\log(t^{-1})A_1 + \left(\log(t^{-1})\|\Sigma\|^{1/2} + 1 \right) A_2 \right. \\ &\quad \left. + \left(1 + \log(t^{-1}) \sum_{i=1}^d \mathbb{E}|W_i| + a \right) A^3 A_3 + aA \right], \end{aligned}$$

where

$$\begin{aligned} A_1 &= \sum_{i,j=1}^d |(\Lambda^{-1})_{j,i}| \sqrt{\mathbb{V}[\mathbb{E}[(W'_i - W_i)(W'_j - W_j) \mid W]]}, \\ A_2 &= \sum_{i,j=1}^d |(\Lambda^{-1})_{j,i}| \sqrt{\mathbb{E}[R_i^2]}, \quad A_3 = \sum_{i=1}^d \max_{j \in \{1, \dots, d\}} |(\Lambda^{-1})_{j,i}|, \end{aligned}$$

C denotes a constant that depends on d , $\sqrt{t} = 2CA^3A_3$ and $a > 1$ is taken from the conditions on \mathcal{G} , defined before Theorem 1.5.

3. AUXILIARY RESULTS

The quenched free-energy Φ defined in (1.11) will appear in the regression condition (2.1).

Lemma 3.1.

For Φ defined in (1.11) we obtain

$$\frac{1}{n} \sum_{j=1}^n \xi_j^i \tanh(\langle \lambda, \xi_j \rangle) = \frac{1}{\beta} (\lambda_i - h \delta_{i,l}) + \frac{\partial}{\partial \lambda_i} \Phi(\lambda).$$

Proof. Differentiating with respect to λ_i yields

$$\begin{aligned} \frac{\partial}{\partial \lambda_i} \Phi(\lambda) &= \frac{1}{\beta} (\lambda_i - h \delta_{i,l}) + \frac{1}{n} \sum_{j=1}^n \frac{\sinh(\langle \lambda, \xi_j \rangle)}{\cosh(\langle \lambda, \xi_j \rangle)} \xi_j^i \\ &= \frac{1}{\beta} (\lambda_i - h \delta_{i,l}) + \frac{1}{n} \sum_{j=1}^n \tanh(\langle \lambda, \xi_j \rangle) \xi_j^i. \end{aligned}$$

Rearranging the equality yields the result. \square

Moreover we consider

$$C_l^n(\xi) := -D^2 \Phi(\lambda_l^n(\xi)) = \frac{1}{\beta} \text{Id}_{\mathbb{R}^p} - \frac{1}{n} \sum_{i=1}^n \cosh^{-2}(\langle \lambda_l^n(\xi), \xi_i \rangle) \xi_i \xi_i^t,$$

with $\lambda_l^n(\xi)$ are defined in Proposition 1.1.

Lemma 3.2.

Let $\beta > 0$ and $h \geq 0$ such that $(\beta, h) \neq (\beta_c, 0)$. Choose an $l \in \mathbb{Z}$, $l \neq 0$, satisfying $|l| \leq p$ in the case of bounded p . Then there exists a constant $c_3 > 0$ such that

$$\sup_{l \in L} \left\| C_l^n(\xi) - \frac{1}{\beta} [1 - \beta(1 - (x^*)^2)] \text{Id}_{\mathbb{R}^p} \right\| \leq c_3 \sqrt{p} \epsilon_n$$

for \mathbb{P}_ξ -almost all ξ and all $n \geq n_1(\xi)$.

Here $\|\cdot\|$ denotes the operator norm. The proof of Lemma 3.2 is given in [12, Lemma 3.2] and uses (1.9), Proposition 3.1 and that with (1.6) x^* satisfies $\cosh^{-2} \operatorname{arctanh} x^* = 1 - (x^*)^2$.

Using the notation

$$m_i^j(\sigma, \xi) := \frac{1}{n} \sum_{\mu=1}^p \sum_{\substack{r=1 \\ r \neq j}}^n \xi_i^\mu \xi_r^\mu \sigma_r, \quad (3.1)$$

$$m_i(\sigma, \xi) := \frac{1}{n} \sum_{\mu=1}^p \sum_{r=1}^n \xi_i^\mu \xi_r^\mu \sigma_r \quad (3.2)$$

the next lemma states an exact expression for the conditional probability that will occur in the linear regression condition (2.1).

Lemma 3.3.

Let $\sigma_i \in \{-1, 1\}$. Then we obtain for the conditional distribution of a single spin

$$P_{n, \beta, h e_l, \xi}(\sigma_i = t \mid (\sigma_k)_{k \neq i}) = \frac{\exp(\beta m_i^i(\sigma, \xi) t + h \xi_i^l t)}{\sum_{k \in \{-1, 1\}} \exp(\beta m_i^i(\sigma, \xi) k + h \xi_i^l k)}$$

and thus

$$\mathbb{E}[\sigma_i \mid (\sigma_k)_{k \neq i}] = \tanh(\beta m_i^i(\sigma, \xi) + h \xi_i^l),$$

where \mathbb{E} denotes the expectation with respect to $P_{n,\beta,he_l,\xi}$.

Proof. Direct calculations yield

$$\begin{aligned} P_{n,\beta,he_l,\xi}(\sigma_i = t \mid (\sigma_k)_{k \neq i}) &= \frac{P_{n,\beta,he_l,\xi}(\{\sigma_i = t\} \cap (\sigma_k)_{k \neq i})}{P_{n,\beta,he_l,\xi}((\sigma_k)_{k \neq i})} \\ &= \frac{\exp \left[\frac{\beta}{2n} \sum_{\mu=1}^p (\xi_i^\mu)^2 + \frac{2\beta}{2n} \sum_{\mu=1}^p \sum_{\substack{j=1 \\ j \neq i}}^n \xi_i^\mu \xi_j^\mu \sigma_j t + \frac{\beta}{2n} \sum_{\mu=1}^p \sum_{\substack{k,j=1 \\ k,j \neq i}}^n \xi_k^\mu \xi_j^\mu \sigma_j \sigma_k + h \xi_i^l t + h \sum_{\substack{j=1 \\ j \neq i}}^n \xi_j^l \sigma_j \right]}{\sum_{k \in \{-1,1\}} \exp \left[\frac{\beta p}{2n} + \frac{2\beta}{2n} \sum_{\mu=1}^p \sum_{\substack{j=1 \\ j \neq i}}^n \xi_i^\mu \xi_j^\mu \sigma_j k + \frac{\beta}{2n} \sum_{\mu=1}^p \sum_{\substack{k,j=1 \\ k,j \neq i}}^n \xi_k^\mu \xi_j^\mu \sigma_j \sigma_k + h \xi_i^l k + h \sum_{\substack{j=1 \\ j \neq i}}^n \xi_j^l \sigma_j \right]} \\ &= \frac{\exp(\beta m_i^i(\sigma, \xi) t + h \xi_i^l t)}{\sum_{k \in \{-1,1\}} \exp(\beta m_i^i(\sigma, \xi) k + h \xi_i^l k)}, \end{aligned}$$

where we canceled equivalent expressions in numerator and denominator and used the expression for $m_i^i(\sigma, \xi)$. Thus

$$\begin{aligned} \mathbb{E}[\sigma_i \mid (\sigma_k)_{k \neq i}] &= P(\{\sigma_i = 1\} \cup (\sigma_k)_{k \neq i}) - P(\{\sigma_i = -1\} \cup (\sigma_k)_{k \neq i}) \\ &= \frac{\exp(\beta m_i^i(\sigma, \xi) + h \xi_i^l) - \exp(-\beta m_i^i(\sigma, \xi) - h \xi_i^l)}{\exp(\beta m_i^i(\sigma, \xi) + h \xi_i^l) + \exp(-\beta m_i^i(\sigma, \xi) - h \xi_i^l)} \\ &= \tanh(\beta m_i^i(\sigma, \xi) + h \xi_i^l). \end{aligned}$$

□

Higher order moments of the rescaled empirical spin vector of the Hopfield model, appearing in Theorems 1.3 up to 1.7, can be bounded as follows:

Lemma 3.4.

For W as in Theorems 1.3 up to 1.7 we obtain that for any $l \in \mathbb{N}$ and $j \in \{1, \dots, p\}$

$$\mathbb{E}|W_j^l| \leq \text{const.}(l).$$

Proof. First we will have to make a transformation with the well-known Hubbard-Stratonovich approach, expressing the distribution of S_n in the Hopfield model in terms of Φ . This approach was for example used in [4, Lemma 2.2] and in [7]. Let Id denote the $p \times p$ identity matrix and for $\beta > 0$ and $h \geq 0$ we pick a random vector V in a way that $\mathcal{L}(V)$ equals a p -dimensional centered Gaussian vector with covariance matrix $\beta^{-1} \text{Id}$ and V is chosen to be independent from all other random variables involved. Additionally $\lambda := \lambda_l^n(\xi)$ denotes the maximum point of Φ taken from Proposition 1.1. First we note that

$$P_{n,\beta,he_l,\xi}(S_n \in dy) = Z_{n,\beta,he_l,\xi}^{-1} \exp \left(\frac{\beta}{2n} \langle y, y \rangle + \langle y, he_l \rangle \right) P_n(S_n \in dy),$$

where $P_n(S_n \in dy) = \prod_{i=1}^n \rho(d\sigma_i)$ and $\rho(d\sigma_i) = \frac{1}{2}\delta_{-1}(d\sigma_i) + \frac{1}{2}\delta_1(d\sigma_i)$. Furthermore for $u \in \mathbb{R}^p$ we have

$$\begin{aligned} \int_{\mathbb{R}^p} \exp\left(\frac{\beta}{n}\langle u, y \rangle + \langle y, he_l \rangle\right) P_n(S_n \in dy) &= \int_{\mathbb{R}^p} \exp\left(\frac{\beta}{n} \sum_{\mu=1}^p \sum_{j=1}^n \xi_j^\mu \sigma_j u_\mu + \sum_{\mu=1}^p \sum_{j=1}^n \xi_j^\mu \sigma_j he_l^\mu\right) \prod_{i=1}^n \rho(d\sigma_i) \\ &= \prod_{i=1}^n \int_{\mathbb{R}} \exp\left(\frac{\beta}{n}\langle \xi_i \sigma_i, u \rangle + \langle \xi_i \sigma_i, he_l \rangle\right) \rho(d\sigma_i) = \exp\left(\sum_{i=1}^n \log \cosh\langle \xi_i, \frac{\beta u}{n} + he_l \rangle\right). \end{aligned}$$

Hence, for $t \in \mathbb{R}$, $x := x_l^n(\xi)$ and $A(n) = \sqrt{n}t + nx$ we obtain

$$\begin{aligned} P\left(V + \sqrt{n}\left(\frac{S_n}{n} - x\right) \leq t\right) &= P(\sqrt{n}V + S_n \leq A(n)) \\ &= Z_{n,\beta,he_l,\xi}^{-1} \int_{\mathbb{R}^p} \exp\left(\frac{\beta}{2n}\langle y, y \rangle + \langle y, he_l \rangle\right) \\ &\quad \cdot \int_{v \leq A(n)-y} \left(\frac{\beta}{2\pi n}\right)^{p/2} \exp\left(-\frac{\beta}{2n}\langle v, v \rangle\right) dv P_n(S_n \in dy) \end{aligned}$$

The substitution $u = v + y$ and abbreviating $C_{p,n} := Z_{n,\beta,he_l,\xi}^{-1} \left(\frac{\beta}{2\pi n}\right)^{p/2}$ yields

$$\begin{aligned} P(V + \sqrt{n}\left(\frac{S_n}{n} - x\right) \leq t) &= C_{p,n} \int_{\mathbb{R}^p} \exp(\langle y, he_l \rangle) \int_{u \leq A(n)} \exp\left(-\frac{\beta}{2n}\langle u, u \rangle\right) \exp\left(\frac{\beta}{n}\langle u, y \rangle\right) du P_n(S_n \in dy). \end{aligned}$$

The abbreviation $\tilde{C}_{p,n} = C_{p,n} n^{p/2}$ yields

$$\begin{aligned} P\left(V + \sqrt{n}\left(\frac{S_n}{n} - x\right) \leq t\right) &= C_{p,n} \int_{u \leq A(n)} \exp\left(-\frac{\beta}{2n}\langle u, u \rangle + \sum_{i=1}^n \log \cosh\langle \xi_i, \frac{\beta u}{n} + he_l \rangle\right) du \\ &= \tilde{C}_{p,n} \int_{z \leq t} \exp\left(-\frac{\beta}{2n}\langle \sqrt{n}z + n\lambda - nhe_l, \sqrt{n}z + n\lambda - nhe_l \rangle\right. \\ &\quad \left.+ \sum_{i=1}^n \log \cosh\langle \xi_i, \frac{\beta z}{\sqrt{n}} + \lambda - he_l + he_l \rangle\right) dz \\ &= \tilde{C}_{p,n} \int_{z \leq t} \exp\left(n\Phi\left(\frac{\beta z}{\sqrt{n}} + \lambda\right)\right) dz, \end{aligned}$$

where we used the substitution $u = \sqrt{n}z + nx$ for the second equality. Thus, we have

$$\mathcal{L}\left(V + \sqrt{n}\left(\frac{S_n}{n} - x\right)\right) = \tilde{Z}_{n,\beta,he_l,\xi}^{-1} \exp\left[n\Phi\left(\lambda + \frac{\beta x}{n}\right)\right] dx, \quad (3.3)$$

where $\tilde{Z}_{n,\beta,he_l,\xi}^{-1}$ denotes a normalization. Applying this transformation does not change the finiteness of any of the moments of the W_j . Thus the new measure has the density (3.3). Using second-order multivariate Taylor expansion of Φ (see (5.1)) and the fact that λ is a maximum point of Φ we see that the density of this new measure with respect to the Lebesgue measure is given by

$$\text{const. exp} \left[-\frac{1}{2} \langle y, -D^2\Phi(\lambda) y \rangle \right]$$

(up to negligible terms). With Proposition 1.1 (a) we know that for any $(\beta, h) \neq (\beta_c, 0)$ the Hessian $-D^2\Phi(\lambda)$ is uniformly positive definite. This fact combined with the transformation of integrals yields that a measure with this density has moments of any finite order. \square

4. PROOFS OF THE THEOREMS

Constructing an exchangeable pair in the Hopfield model to obtain an approximate linear regression property (2.1) leads us to Φ taken from (1.11). Let $(\beta, h) \neq (\beta_c, 0)$, and let $x := x_l^n(\xi)$ denote the unique global maximum point of Φ , see Proposition 1.1. For $k \in \mathbb{N}$ fixed, $k \leq p$, we consider

$$W := \sqrt{n} \pi_k \left(\frac{S_n}{n} - x \right) = \sqrt{n} \left(\frac{1}{n} \sum_{j=1}^n \xi_j^1 \sigma_j - x_1, \dots, \frac{1}{n} \sum_{j=1}^n \xi_j^k \sigma_j - x_k \right)^t.$$

We start by constructing an exchangeable pair. Therefore we produce a spin collection $\sigma' = (\sigma'_i)_{i \geq 1}$ via a *Gibbs sampling procedure*: We take I to be a random variable that is uniformly distributed over $\{1, \dots, n\}$ and independent from all other random variables involved. Exchanging the spin σ_i with σ'_i drawn from the conditional distribution of the i^{th} coordinate given $(\sigma_j)_{j \neq i}$ under $P_{n,\beta,he_l,\xi}$, independently from σ_i , we obtain

$$W' := W + \frac{1}{\sqrt{n}} (\xi_I^1 \sigma'_I, \dots, \xi_I^k \sigma'_I) - \frac{1}{\sqrt{n}} (\xi_I^1 \sigma_I, \dots, \xi_I^k \sigma_I). \quad (4.1)$$

In this case (W, W') is an exchangeable pair. Let $\mathcal{F} := \sigma(\sigma_i, \xi_j^\mu | i, j, \mu \in \mathbb{N})$. We obtain that for any $i = 1, \dots, k$:

$$\mathbb{E}[W'_i - W_i | \mathcal{F}] = \frac{1}{\sqrt{n}} \mathbb{E} [\xi_I^i \sigma'_I - \xi_I^i \sigma_I | \mathcal{F}].$$

Using the law of total probability for the conditional expectation and independence we have

$$\mathbb{E}[W'_i - W_i | \mathcal{F}] = \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \mathbb{E} [\xi_j^i \sigma'_j - \xi_j^i \sigma_j | \mathcal{F}].$$

Since σ_i and ξ_j^i , $i, j \in \mathbb{N}$, are measurable with respect to \mathcal{F} we obtain

$$\mathbb{E}[W'_i - W_i | \mathcal{F}] = -\frac{1}{\sqrt{n}} \frac{1}{n} S_{n,i} + \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \xi_j^i \mathbb{E} [\sigma'_j | \mathcal{F}].$$

With the help of independence and the construction of the exchangeable pair we obtain $\mathbb{E}[\sigma'_j | \mathcal{F}] = \mathbb{E}[\sigma'_j | \sigma_1, \dots, \sigma_n] = \mathbb{E}[\sigma_j | (\sigma_k)_{k \neq j}]$. Applying Lemma 3.3 yields

$$\begin{aligned} \mathbb{E}[W'_i - W_i | \mathcal{F}] &= -\frac{1}{\sqrt{n}} \frac{1}{n} S_{n,i} + \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \xi_j^i \tanh(\beta m_j^j(\sigma, \xi) + h \xi_j^l) \\ &= -\frac{1}{\sqrt{n}} \frac{1}{n} S_{n,i} + \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \xi_j^i \tanh(\beta m_j(\sigma, \xi) + h \xi_j^l) + R_{1,i}, \end{aligned}$$

with

$$R_{1,i} := \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \xi_j^i \left[\tanh(\beta m_j^j(\sigma, \xi) + h \xi_j^l) - \tanh(\beta m_j(\sigma, \xi) + h \xi_j^l) \right]. \quad (4.2)$$

Now it is important to note that

$$\tanh(\beta m_j(\sigma, \xi) + h \xi_j^l) = \tanh\left\langle \beta \frac{S_n}{n} + h e_l, \xi_j \right\rangle.$$

Thus, with Lemma 3.1, we have

$$\frac{1}{n} \sum_{j=1}^n \xi_j^i \tanh(\beta m_j(\sigma, \xi) + h \xi_j^l) = \frac{1}{\beta} \left(\beta \frac{S_{n,i}}{n} + h \delta_{i,l} - h \delta_{i,l} \right) + \frac{\partial}{\partial \lambda_i} \Phi \left(\beta \frac{S_n}{n} + h e_l \right).$$

This equation yields

$$\mathbb{E}[W'_i - W_i | \mathcal{F}] = \frac{1}{\sqrt{n}} \frac{\partial}{\partial \lambda_i} \Phi \left(\beta \frac{S_n}{n} + h e_l \right) + R_{1,i}. \quad (4.3)$$

We continue by applying (1.12) and (5.2) (see Appendix) to the first summand in (4.3). Since $\lambda_l^n(\xi)$ is a unique maximum point of $\Phi(\lambda)$ we have $\frac{\partial}{\partial \lambda_i} \Phi(\lambda_l^n(\xi)) = 0$. We also note that $\frac{\beta S_{n,i}}{n} + h \delta_{i,l} - (\lambda_l^n(\xi))_i = \beta \frac{W_i}{\sqrt{n}}$. Thus, the first summand in (4.3) is equal to

$$\frac{1}{\sqrt{n}} \sum_{t=1}^k \left(\frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right) \frac{\beta W_t}{\sqrt{n}} + R_{2,i},$$

with

$$R_{2,i} := \sum_{t=k}^p \left(\frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right) \frac{\beta W_t}{n} + \sum_{l,t=1}^p \mathcal{O} \left(\frac{1}{\sqrt{n}} \frac{W_l}{\sqrt{n}} \frac{W_t}{\sqrt{n}} \right). \quad (4.4)$$

Abbreviating

$$R(i) := R_{1,i} + R_{2,i}, \quad (4.5)$$

we have

$$\begin{aligned} \mathbb{E}[W'_i - W_i | \mathcal{F}] &= \frac{1}{n} \sum_{t=1}^k \left(\frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right) \beta W_t + R(i) \\ &= \frac{\beta}{n} \langle [D^2 \Phi(\lambda_l^n(\xi))]_{i,k}, W \rangle + R(i), \end{aligned} \quad (4.6)$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product and $[D^2 \Phi(\lambda_l^n(\xi))]_{i,k}$ denotes the first k entries of the i^{th} row of the matrix $D^2 \Phi(\lambda_l^n(\xi))$. We obtain

$$\mathbb{E}[W' - W | \mathcal{F}] = \frac{\beta}{n} [D^2 \Phi(\lambda_l^n(\xi))]_{|k \times k} W + R(W), \quad (4.7)$$

with $R(W) = (R(1), \dots, R(k))$. We define $\Lambda := \frac{\beta}{n} \left[-D^2 \Phi(\lambda_l^n(\xi)) \right]_{|k \times k}$. With Proposition 1.1(a) $-D^2 \Phi(\lambda_l^n(\xi))$ is uniformly positive definite and thus Λ is invertible. We conducted the linear regression condition for the sigma-algebra \mathcal{F} but it should be noted that it yields also the linear regression condition for the sigma-algebra generated by W since W is measurable with respect to \mathcal{F} . In this case the linear regression condition (2.1) is fulfilled.

Proof of Theorem 1.3. With (4.7) we are able to apply Theorem 2.1. Since the Hessian matrix of Φ and β itself are constants we have $\lambda^{(i)} = \mathcal{O}(n)$. We continue by estimating C taken from Theorem 2.1. We start by giving a bound for $R_{1,i}$, defined in (4.2). Since the $\tanh(x)$ is 1-Lipschitz we obtain

$$\begin{aligned} |R_{1,i}| &= \left| \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \xi_j^i \left[\tanh(\beta m_j^i(\sigma, \xi) + h \xi_j^l) - \tanh(\beta m_j(\sigma, \xi) + h \xi_j^l) \right] \right| \\ &\leq \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \left| \beta m_j^i(\sigma, \xi) + h \xi_j^l - (\beta m_j(\sigma, \xi) + h \xi_j^l) \right| \\ &= \frac{\beta}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \left| \frac{1}{n} \sum_{\mu=1}^p (\xi_j^\mu)^2 \sigma_j \right| \\ &= \frac{\beta}{\sqrt{n}} \frac{1}{n} \sum_{j=1}^n \left| \frac{1}{n} p \sigma_j \right| \leq \frac{\beta p}{\sqrt{n}} \frac{1}{n}. \end{aligned}$$

For the estimation of $R_{2,i}$ we note that by Lemma 3.4 we have for the second part of (4.4)

$$\mathbb{E} \left[\sum_{l,t=1}^p \mathcal{O} \left(\frac{1}{\sqrt{n}} \frac{W_l}{\sqrt{n}} \frac{W_t}{\sqrt{n}} \right) \right] = \mathcal{O} \left[\frac{p^2}{n^{3/2}} \right].$$

For the first part of (4.4) we note that by Lemma 3.2, since $i \notin \{k+1, \dots, p\}$ and $t \in \{1, \dots, k\}$,

$$\left| \frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right| \leq c_3 \sqrt{p} \epsilon_n$$

since this expression is a non-diagonal entry of the matrix $-C_l^n(\xi)$. Thus we obtain that

$$\mathbb{E} \left[\sum_{t=k}^p \left(\frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right) \frac{\beta W_t}{n} \right] = \mathcal{O} \left[\frac{p \sqrt{p} \epsilon_n}{n} \right],$$

and finally

$$\mathbb{E} |R_{2,i}| = \mathcal{O} \left[\max \left\{ \frac{p \sqrt{p} \epsilon_n}{n}, \frac{p^2}{n^{3/2}} \right\} \right]. \quad (4.8)$$

Thus we have

$$C = \sum_{i=1}^k \lambda^{(i)} \sqrt{\mathbb{E} [R(i)^2]} = \mathcal{O} \left[\max \left\{ p \sqrt{p} \epsilon_n, \frac{p^2}{n^{1/2}} \right\} \right].$$

The next thing we notice is that for all $i \in \{1, \dots, k\}$

$$|W'_i - W_i| = \left| \frac{1}{\sqrt{n}} \xi_I^i (\sigma'_I - \sigma_I) \right| \leq \frac{1}{\sqrt{n}}.$$

We easily obtain that the bound $B = \mathcal{O}(n^{-1/2})$. The only thing left to do is to calculate the tedious conditional variance in A . We have:

$$\begin{aligned} \mathbb{E}[(W'_i - W_i)(W'_j - W_j) \mid \mathcal{F}] &= \frac{1}{n^3} \sum_{t,r=1}^n \xi_t^i \sigma_t \xi_r^j \sigma_r + \frac{1}{n^3} \sum_{t,r=1}^n \mathbb{E}[\xi_t^i \sigma'_t \xi_r^j \sigma'_r \mid \mathcal{F}] \\ &\quad - \frac{2}{n^3} \sum_{t,r=1}^n \xi_r^j \xi_t^i \sigma_r \mathbb{E}[\sigma'_t \mid \mathcal{F}] \\ &=: A_1 + A_2 + A_3. \end{aligned} \tag{4.9}$$

To bound the variances of these three terms we abbreviate

$$\tilde{m}_i(\sigma, \xi) := \frac{1}{n} \sum_{t=1}^n \xi_t^i \sigma_t = \frac{1}{\sqrt{n}} W_i + x_i.$$

Thus,

$$\begin{aligned} \mathbb{V}[A_1] &= \frac{1}{n^2} \mathbb{V}[\tilde{m}_i(\sigma) \tilde{m}_j(\sigma)] = \frac{1}{n^2} \mathbb{V}\left[\frac{W_i W_j}{n} + \frac{W_i}{\sqrt{n}} x_j + \frac{W_j}{\sqrt{n}} x_i\right] \\ &\leq \frac{1}{n^2} \text{const.} \max\left\{\frac{1}{n^2} \mathbb{V}[W_i W_j], \frac{1}{n} \mathbb{V}[W_i]\right\} \\ &\leq \frac{1}{n^2} \frac{\text{const.}}{n^2} (\mathbb{E}[W_i^2 W_j^2] + n \mathbb{E}[W_i]). \end{aligned}$$

Using Lemma 3.4 we obtain $\mathbb{V}[A_1] = \mathcal{O}(n^{-3})$. For A_2 we obtain

$$A_2 = \frac{1}{n^3} \sum_{t,r=1}^n \mathbb{E}[\xi_t^i \sigma'_t \xi_r^j \sigma'_r \mid \mathcal{F}] = \frac{1}{n} \mathbb{E}\left[\left(\frac{1}{n} \sum_{t=1}^n \xi_t^i \sigma'_t\right) \left(\frac{1}{n} \sum_{r=1}^n \xi_r^j \sigma'_r\right) \mid \mathcal{F}\right].$$

Next we use the identity $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ for a random variable X and a conditional version of Jensen's inequality in order to obtain that $\mathbb{V}[A_2] \leq \mathbb{V}[A_1] = \mathcal{O}(n^{-3})$, since σ' is an identical copy of σ . With Lemma 3.3 we get

$$\begin{aligned} -A_3/2 &= \frac{1}{n^3} \sum_{t,r=1}^n \xi_r^j \sigma_r \mathbb{E}[\xi_t^i \sigma'_t \mid \mathcal{F}] \\ &= \frac{1}{n^3} \sum_{t,r=1}^n \xi_r^j \sigma_r \xi_t^i \tanh(m_t^t(\sigma, \xi) + h \xi_t^l) \\ &= \frac{1}{n^3} \sum_{t,r=1}^n \xi_r^j \sigma_r \xi_t^i [\tanh(m_t^t(\sigma, \xi) + h \xi_t^l) - \tanh(m_t(\sigma, \xi) + h \xi_t^l)] \\ &\quad + \frac{1}{n^3} \sum_{t,r=1}^n \xi_r^j \sigma_r \xi_t^i \tanh(m_t(\sigma, \xi) + h \xi_t^l) \\ &=: M_1 + M_2. \end{aligned} \tag{4.10}$$

Using the same estimations as for $R_n^{(1)}(i)$ we obtain

$$M_1 \leq \left| \frac{1}{n^2} \sum_{r=1}^n \xi_r^j \sigma_r \right| \left| \frac{\beta p}{n} \right| = \left| \frac{1}{n} \beta p \left(\frac{W_j}{\sqrt{n}} + x_j \right) \right|.$$

Hence $\mathbb{V}[M_1] = \mathcal{O}\left[\frac{p^2}{n^3}\right]$ by Lemma 3.4. Additionally we get by using Lemma 3.1, (5.2) and the abbreviation $\Phi^{(2),i,j}(\lambda) := \frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi))$

$$\begin{aligned} M_2 &= \frac{1}{n} \left(\frac{W_j}{\sqrt{n}} + x_j \right) \left(\frac{W_i}{\sqrt{n}} + x_i + \frac{\partial}{\partial \lambda_i} \Phi \left(\beta \frac{S_n}{n} + h e_l \right) \right) \\ &= \left(\frac{W_j}{n\sqrt{n}} + \frac{x_j}{n} \right) \left(\frac{W_i}{\sqrt{n}} + x_i + \sum_{t=1}^p \left(\Phi^{(2),i,t}(\lambda) \right) \frac{\beta W_t}{\sqrt{n}} + \sum_{l,t=1}^p \mathcal{O} \left[\frac{W_l W_t}{n} \right] \right). \end{aligned}$$

Since we are estimating the variance of the expressions, constant expressions will vanish. Hence using Lemma 3.4 and Lemma 3.2 in the same way as for (4.8) we have

$$\mathbb{V}[M_2] = \mathcal{O} \left[\max \left\{ \frac{p^3 \epsilon_n^2}{n^3}, \frac{p^2}{n^3} \right\} \right].$$

Therefore $\mathbb{V}[A_3]$ can be bounded by $\mathcal{O} \left[\max \left\{ \frac{p^3 \epsilon_n^2}{n^3}, \frac{p^2}{n^3} \right\} \right]$. Thus the variance in A of Theorem 2.1 can be bounded by 9 times the maximum of the variances of A_1, A_2, A_3 . Consequently we obtain

$$A = \sum_{i,j=1}^k \lambda^{(i)} \sqrt{\mathbb{V} \left[\mathbb{E}[(W'_i - W_i)(W'_j - W_j) | W] \right]} = \mathcal{O} \left[\max \left\{ \frac{p^{3/2} \epsilon_n}{n^{1/2}}, \frac{p}{\sqrt{n}} \right\} \right]$$

and this completes the proof. \square

Proof of Theorem 1.5. Having seen the proof of Theorem 1.3 this proof gets very simple. We first note that Theorem 2.2 can be applied since the regression condition is the same as for Theorem 1.3. A_1 matches A taken from the same proof and thus $\log(n)A_1 = \mathcal{O} \left[\log(n) \max \left\{ \frac{p^{3/2} \epsilon_n}{n^{1/2}}, \frac{p}{n} \right\} \right]$. Using Lemma 3.4 and the estimation of the C-term in 1.3 we have that the second expression is $\mathcal{O} \left[\log(n) \max \left\{ \frac{p^{3/2} \epsilon_n}{n^{1/2}}, \frac{p}{\sqrt{n}} \right\} \right]$. The same Lemma, $A = \frac{1}{\sqrt{n}}$ and $A_3 = \mathcal{O}(n)$ yield that the third and fourth expression have the order $\mathcal{O}(\log(n)n^{-1/2})$. Thus the theorem is proven. \square

Proof of Theorem 1.6. In order to prove the theorem we have to make small adjustments to the proof of Theorem 1.3. Using the same techniques as before we arrive at

$$\mathbb{E}[W' - W | \mathcal{F}] = \frac{\beta}{n} \left[D^2 \Phi(\lambda_l^n(\xi)) \right] W + R(W),$$

with $R(W) = (R(1), \dots, R(p))$, where $R(i) = R_{1,i} + \tilde{R}_{2,i}$ with $R_{1,i}$ taken from (4.2) and

$$\tilde{R}_{2,i} := \sum_{l,t=1}^p \mathcal{O} \left(\frac{1}{\sqrt{n}} \frac{W_l}{\sqrt{n}} \frac{W_t}{\sqrt{n}} \right). \quad (4.11)$$

This expression is the central difference to the proof of Theorem 1.3. Whereas the expression (4.4) contained the expression

$$\sum_{t=k}^p \left(\frac{\partial^2}{\partial \lambda_i \partial \lambda_t} \Phi(\lambda_l^n(\xi)) \right) \frac{\beta W_t}{n}, \quad (4.12)$$

which made us use Lemma 3.2, (4.12) is now part of ΛW since p is a constant and we do not need a projection to define W . Thus our expression (4.11) contains just the second expression of the right hand side of (4.4). Fortunately this can be estimated using Lemma 3.4. Thus, without

using Lemma 3.2, the computation of the rate of convergence gets a lot easier. Again it only remains to estimate A , B and C taken from Theorem 2.1. We note that B is the same as in Theorem 1.3. Thus $B = \mathcal{O}(n^{-1/2})$. $R_{1,i}$ is the same as in (4.2) and is bounded in the same way as in Theorem 1.3. Since $\tilde{R}_{2,i}$ was part of (4.4) and p is fixed we obtain by using Lemma 3.4

$$\mathbb{E}|\tilde{R}_{2,i}| = \mathcal{O}(n^{-3/2}). \quad (4.13)$$

In comparison to Theorem 1.3 and the bound in (4.8) we notice that the first part of the maximum is not existent since the expression (4.12) is not part of $\tilde{R}_{2,i}$ and the second part of the maximum is the same as the bound in (4.13) with p constant. Using the bound on $R_{1,i}$ and $\tilde{R}_{2,i}$ we obtain $C = \mathcal{O}(n^{-1/2})$. If we split the expectation of the expression A in the same way as in (4.9) and we note that A_1 and A_2 are estimated in exact the same way as for the proof of Theorem 1.3. Finally we note that for p fixed we can also split A_3 as in (4.10) and that with the same reasons that led to (4.13) $\mathbb{V}[M_1] = \mathbb{V}[M_2] = \mathcal{O}(n^{-3})$. Hence, $A = \mathcal{O}(n^{-1/2})$. \square

Proof of Theorem 1.7. The proof uses the fact that the conditional joint distribution of the $(\sigma_i)_i$, conditioned on the event $\left\{ \left\| \frac{S_n}{n} - x^* e_l \right\| < \epsilon \right\}$, is given by

$$P_{n,\beta,\xi}(\sigma) = \frac{1}{\tilde{Z}_{n,\beta,\xi}} \exp\left(-\beta H_n(\sigma, \xi)\right) \mathbf{1}_{B(x^* e_l, \epsilon)}\left(\frac{S_n}{n}\right),$$

where $\tilde{Z}_{n,\beta,\xi}$ denotes a normalization. Thus we are able to follow the lines of the proof of Theorem 1.3. \square

5. APPENDIX

For the proofs of the theorems for the Hopfield model we need a multivariate *second-order Taylor expansion* of $\Phi(\lambda)$ defined in (3.1). Let us denote by $D^2\Phi(\lambda)$ the Hessian matrix $\{\partial^2\Phi(\lambda)/\partial\lambda_i\partial\lambda_j, i, j = 1, \dots, p\}$ of Φ at λ . We obtain

$$\begin{aligned} \Phi(u) &= \Phi(\lambda) + \sum_{k=1}^p \frac{\partial}{\partial u_k} \Phi(\lambda) (u_k - \lambda_k) + \frac{1}{2} \langle (u - \lambda), D^2\Phi(\lambda) \cdot (u - \lambda) \rangle \\ &\quad + \frac{1}{6} \sum_{t,k,j=1}^p \tilde{R}_{t,k,j} (u_t - \lambda_t) (u_k - \lambda_k) (u_j - \lambda_j), \end{aligned} \quad (5.1)$$

with $|\tilde{R}_{t,k,j}| \leq \left\| \frac{\partial^3}{\partial u_k \partial u_t \partial u_j} \Phi \right\|$. For any fixed $m \in \{1, \dots, p\}$ and any $\lambda, u \in \mathbb{R}^p$ it follows that

$$\begin{aligned} \frac{\partial}{\partial u_m} \Phi(u) &= \frac{\partial}{\partial u_m} \Phi(\lambda) + \sum_{k=1}^p \frac{\partial^2}{\partial u_k \partial u_m} \Phi(\lambda) (u_k - \lambda_k) \\ &\quad + \sum_{k,t=1}^p \mathcal{O}((u_k - \lambda_k)(u_t - \lambda_t)). \end{aligned} \quad (5.2)$$

REFERENCES

- [1] Daniel J. Amit, Hanoach Gutfreund, and H. Sompolinsky, *Spin-glass models of neural networks*, Phys. Rev. A (3) **32** (1985), no. 2, 1007–1018. MR 797031 (86g:92015)
- [2] A. Bovier and V. Gayraud, *An almost sure large deviation principle for the Hopfield model*, Ann. Probab. **24** (1996), no. 3, 1444–1475.

- [3] ———, *The retrieval phase of the Hopfield model: a rigorous analysis of the overlap distribution*, Probab. Theory Related Fields **107** (1997), no. 1, 61–98.
- [4] A. Bovier, V. Gayraud, and P. Picco, *Gibbs states of the Hopfield model in the regime of perfect memory*, Probab. Theory Related Fields **100** (1994), no. 3, 329–363.
- [5] S. Chatterjee, P. Diaconis, and E. Meckes, *Exchangeable pairs and Poisson approximation*, Probab. Surv. **2** (2005), 64–106 (electronic).
- [6] S. Chatterjee and Q.-M. Shao, *Stein’s method of exchangeable pairs with application to the Curie-Weiss model*, to appear in Ann. Appl. Prob., 2010.
- [7] P. Eichelsbacher and M. Löwe, *Moderate deviations for the overlap parameter in the Hopfield model*, Probab. Theory and Related Fields **130** (2004), no. 4, 441–472.
- [8] P. Eichelsbacher and M. Löwe, *Stein’s-method for dependent variabels occurring in statistical mechanics*, Electron. J. Probab. **15** (2010), no. 30, 962–988.
- [9] P. Eichelsbacher and B. Martschink, *On rates of convergence in the curie-weiss-potts model with external field*, preprint, arXiv:1011.0319v1, 2013.
- [10] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, New York, 1985.
- [11] B. Gentz, *An almost sure central limit theorem for the overlap parameters in the Hopfield model*, Stochastic Process. Appl. **62** (1996), no. 2, 243–262.
- [12] ———, *A central limit theorem for the overlap in the Hopfield model*, Ann. Probab. **24** (1996), no. 4, 1809–1841.
- [13] B. Gentz and M. Löwe, *Fluctuations in the Hopfield model at the critical temperature*, Markov Process. Related Fields **5** (1999), no. 4, 423–449.
- [14] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. U.S.A. **79** (1982), 2554–2558.
- [15] L. A. Pastur and A. L. Figotin, *Exactly soluble model of a spin glass*, Sov. J. Low Temp. Phys. **3** (1977), no. 6, 378–383.
- [16] ———, *On the theory of disordered spin systems*, Theor. Math. Phys. **35** (1977), 403–414.
- [17] G. Reinert and A. Röllin, *Multivariate normal approximation with Stein’s method of exchangeable pairs under a general linearity condition*, Ann. Probab. **37** (2009), no. 6, 2150–2173.
- [18] Y. Rinott and V. Rotar, *On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics*, Ann. Appl. Probab. **7** (1997), no. 4, 1080–1105.
- [19] Yosef Rinott and Vladimir Rotar, *A multivariate CLT for local dependence with $n^{-1/2} \log n$ rate and applications to multivariate graph related statistics*, J. Multivariate Anal. **56** (1996), no. 2, 333–350. MR 1379533 (97a:60035)
- [20] C. Stein, *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory (Berkeley, Calif.), Univ. California Press, 1972, pp. 583–602. MR MR0402873 (53 #6687)
- [21] ———, *Approximate computation of expectations*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7, Institute of Mathematical Statistics, Hayward, CA, 1986. MR MR882007 (88j:60055)
- [22] C. Stein, P. Diaconis, S. Holmes, and G. Reinert, *Use of exchangeable pairs in the analysis of simulations, Stein’s method: expository lectures and applications*, IMS Lecture Notes Monogr. Ser., vol. 46, Inst. Math. Statist., Beachwood, OH, 2004, pp. 1–26. MR MR2118600 (2005j:65005)
- [23] M. Talagrand, *On the Hopfield model at the critical temperature*, Probab. Theory Related Fields **121** (2001), no. 2, 237–268.